LECTURE 6 : PREDICTIVE DISTRIBUTION FOR CLASSIFICATION

Goal: To obtain $p(t_* | \bar{t})$

Meaning, given $\bar{x}_*$ and $\{\bar{x}_n, t_n\}_{n=1}^N$, where $t_n \in \{0, 1\}$ we want to know the label/class of $\bar{x}_*$ $(= t_*)$ and its probability : $p(t_* | \bar{t})$

↓

Knowing $p(t_* | \bar{t})$ will enable us to know t_* (the label of $\bar{x}_*$)

Solution:

$$\underset{(1\times 1)\ (N\times 1)}{p(t_* | \bar{t})} = \int p(t_*, \bar{w} | \bar{t})\, d\bar{w}$$

$$= \int \underbrace{p(t_* | \bar{w})}_{①} \underbrace{p(\bar{w} | \bar{t})}_{②} d\bar{w}$$

① $p(t_* | \bar{w}) = \sigma(\bar{w}^T \bar{\phi}(\bar{x}_*)) = \sigma(\bar{w}^T \bar{\phi}_*)$

② $p(\bar{w} | \bar{t}) = \dfrac{p(\bar{t} | \bar{w})\, p(\bar{w})}{p(\bar{t})}$

We have solved it in the prev. lecture (Full Bayesian)

$$= \frac{\prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}\ G(\bar{w}; \bar{m}_0, S_0)}{\int \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}\ G(\bar{w}; \bar{w}_{MAP}, S_N)\, d\bar{w}}$$

$$\simeq q(\bar{w}) = G(\underset{M\times 1}{\bar{w}}; \underset{M\times 1}{\bar{w}_{MAP}}, \underset{M\times M}{S_N})$$

$$\boxed{p(t_* | \bar{t}) = \int \sigma(\bar{w}^T \bar{\phi}_*)\, G(\bar{w}; \bar{w}_{MAP}, S_N)\, d\bar{w}}$$

Q: How to do the integration?

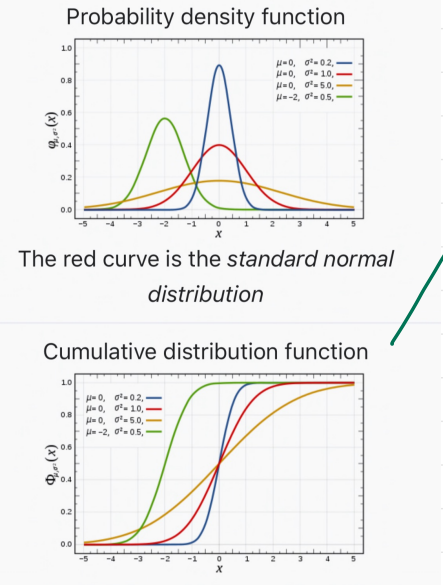
$$p(t_* \mid \bar{t}) = \int \sigma(\bar{w}^T \bar{\phi}_*)\, G(\bar{w}; \bar{w}_{MAP}, S_N)\, d\bar{w}$$

Solution:

Probit function: $\Phi(x) = \int_{-\infty}^{x} G(\theta; 0, 1)\, d\theta$

Cumulative distribution of a Gaussian

The shape is similar to $\sigma(x)$.



Probit convolution:

$$\int \Phi(\lambda x)\, G(x; \mu, \sigma^2)\, dx = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

By assuming $\Phi \approx \sigma$, then probit convolution will solve the integration.

Problem: Probit convolution is an integration over a scalar variable, x; yet our integral function is over a vector, $\bar{w}$.

↓↓

We need to transform our function, so that we can integrate over a scalar value: a_*.

$$a_* = \bar{w}^T \bar{\phi}_* \quad \rightarrow \text{scalar value}$$

$(1\times 1) \quad (1\times M) \quad (M\times 1)$

$$p(t_* \mid \bar{t}) = \int \sigma(a_*)\, G(\bar{w}; \bar{w}_{MAP}, S_N)\, d\bar{w}$$

we need to change them, but how?

Marginalization over a_* (instead of $\bar{w}$)

$a_* = \bar{w}^T \bar{\phi}_*$ ↓ fixed as $\bar{X}_*$ & basis function are fixed

$$p(t_* | \bar{t}) = \int p(t_* | a_*)\, p(a_* | \bar{t})\, da_*$$

$$= \int \sigma(a_*)\, p(a_* | \bar{t})\, da_*$$

Recall: $p(\bar{w} | \bar{t}) \approx q(\bar{w}) = G(\bar{w}; \bar{w}_{MAP}, S_N)$

$p(a_* | \bar{t}) \approx q(a_*) = q(\bar{w}^T \bar{\phi}_*)$

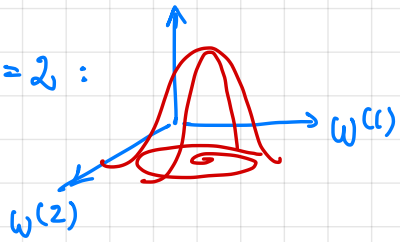
since: $q(\bar{w}) = G(\bar{w}; \bar{w}_{MAP}, S_N)$, $q(a_*)$ should be a Gaussian as the transformation from $q(\bar{w})$ to $q(a_*)$ is a linear transformation:

$G(\bar{w}; \bar{\mu}_w, \Sigma_w)$ (M×1, M×M) → linear transformation → $G(\bar{a}; \bar{\mu}_a, \Sigma_a)$ (R×1, R×R)

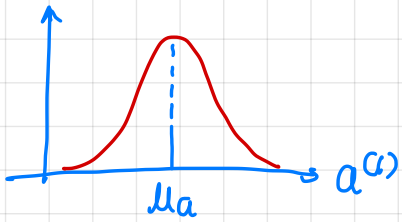
$\bar{a} = A\bar{w} + \bar{b}$

R×1 R×M M×1 R×1

$M = 2$:



$R = 1$:



$\bar{\mu}_a = A\bar{\mu}_w + \bar{b}$; $\Sigma_a = A \Sigma_w A^T$

R×1 R×M M×1 R×1 ; R×R R×M M×M M×R

Hence: $p(a_* | \bar{t}) = G(a_*; a_{MAP}, \Sigma_*)$

where: $a_* = \bar{w}^T \bar{\phi}_* \rightarrow A = \bar{\phi}^T$; $b = 0$

$a_{MAP} = \bar{w}_{MAP}^T \bar{\phi}_*$; $\Sigma_* = \bar{\phi}_*^T S_N \bar{\phi}_*$

1×1 1×M M×1 ; 1×1 1×M M×M M×1

$$p(t_* \mid \bar{t}) = \int \sigma(a_*)\, G(a_*;\, a_{MAP},\, \Sigma_*)\, da_*$$

Recall: Probit convolution:

$$\int \Phi(\lambda x)\, G(x;\, \mu, \sigma^2)\, dx = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

We assume: $\sigma(a_*) = \Phi(\lambda x)$; $x = \dfrac{a_*}{\lambda}$

Thus:

$$p(t_* \mid \bar{t}) = \int \sigma(a_*)\, G\left(\frac{a_*}{\lambda};\, \frac{a_{MAP}}{\lambda},\, \Sigma_*\right) \frac{da_*}{\lambda}$$

$$= \sigma\left(\frac{a_{MAP}/\lambda}{(\lambda^{-2} + \Sigma_*)^{1/2}}\right)$$

Probit function Φ is the same as σ if $\lambda^2 = \dfrac{\pi}{8}$:

$$p(t_* \mid \bar{t}) = \sigma\left(\frac{a_{MAP}}{\sqrt{\frac{\pi}{8}}\left(\left(\frac{\pi}{8}\right)^{-1} + \Sigma_*\right)^{1/2}}\right)$$

Finally:

$$\boxed{p(t_* \mid \bar{t}) = \sigma\left(\frac{a_{MAP}}{\left(1 + \frac{\pi}{8}\Sigma_*\right)^{1/2}}\right)}$$

where: $a_{MAP} = \bar{w}_{MAP}^T \bar{\phi}_*$; $\Sigma_* = \phi_*^T S_N \phi_*$

a_{MAP}: 1×1 ; $\bar{w}_{MAP}^T$: $1\times M$; $\bar{\phi}_*$: $M\times 1$; Σ_*: 1×1 ; ϕ_*^T: $1\times M$; S_N: $M\times M$; ϕ_*: $M\times 1$

Notes:

1. Probit is the cumulative distribution function (cdf) of a Gaussian. When this Gaussian multiplied with another Gaussian, it yields another Gaussian. And, the cdf of the last Gaussian itself is another probit.

$$\int \Phi(\lambda a)\, G(a;\mu,\sigma^2)\, da = \int_{-\infty}^{\lambda a}\int G(\theta;1,0)\, G(a;\mu,\sigma^2)\, da\, d\theta$$

This integration generates a Gaussian

↓

and the integration of this Gaussian (which is the cdf) yields another probit.

MULTICLASS CLASSIFICATION

[•] Binary Logistic Regression : $t_n \in \{0, 1\}$

Training stage : Input : $\{\bar{x}_n, t_n\}_{n=1}^{N}$

Output : $\bar{w}$

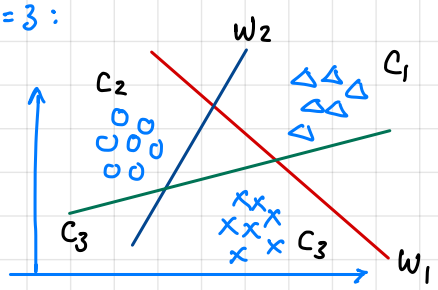
Likelihood : $p(t_n = 1 | \bar{w}) = \sigma(w^T \bar{\phi}(\bar{x}_n)) = \sigma(a_n) = \dfrac{p(\bar{w} | t_n = 1)\, p(t_n = 1)}{\sum_{t_n \in \{0,1\}} p(\bar{w} | t_n = 1)\, p(t_n = 1)}$

We define it in this way

[•] Multiclass Logistic Regression :

$$\underset{K \times 1}{\bar{t}_n} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} C_1 = 1 \\ C_2 = 0 \\ \vdots \\ C_K = 0 \end{bmatrix}$$

$K = 3$:



Output : $\underset{M \times K}{W}$

$t_{n,k} \sim y_{n,k} \leftarrow \bar{w}_k^T \bar{\phi}(\bar{x}_n)$

$\underset{K \times 1}{\bar{t}_n} \sim \underset{K \times 1}{\bar{y}_n} \leftarrow \underset{K \times M}{W^T} \underset{M \times 1}{\bar{\phi}(\bar{x}_n)}$

Likelihood :

$$p(t_n = c_k | W) = \frac{p(W | t_n = c_k)\, p(t_n = c_k)}{p(W)} = \frac{p(W | t_n = c_k)\, p(t_n = c_k)}{\sum_{c_j \in \{c_1 .. c_K\}} p(W | t_n = c_j)\, p(t_n = c_j)}$$

Let's define :

$$\exp(a_{n,k}) = p(W | t_n = c_k)\, p(t_n = c_k)$$

Hence :

$$p(t_n = c_k | W) = \frac{\exp(a_{n,k})}{\sum_{j=1}^{K} \exp(a_{n,j})}$$

$a_{n,k} = \bar{w}_k^T \bar{\phi}(\bar{x}_n)$

Known as Softmax

Implying : $t_{n,k} = 1$: $a_{n,k}$ is encouraged to be large

$t_{n,k} = 0$: $a_{n,k}$ is encouraged to be small

MLE for Multiclass Classification :

(1) Likelihood Data Modeling :

$$\underset{N\times K}{p(\mathbb{T}\mid W)} = \underset{K\times 1}{p(\bar{t}_1, \dots, \bar{t}_N \mid \underset{M\times 1}{\bar{w}_1, \dots, \bar{w}_K})}$$

$$= \prod_n \prod_k p(t_n = c_k \mid \bar{w}_k)^{t_{n,k}}$$

$$= \prod_n \prod_k \left(\frac{\exp(a_{n,k})}{\sum_j \exp(a_{n,j})} \right)^{t_{n,k}}$$

(2) Error Function :

$$E(\bar{w}_1, \dots, \bar{w}_K) = -\log p(\mathbb{T} \mid W)$$

$$= -\sum_n \sum_k t_{nk} \log y_{nk}$$

$$= -\sum_n \sum_k t_{nk} \log \left(\frac{\exp(a_{nk})}{\sum_j \exp(a_{nj})} \right) \quad ; \; a_{nk} = \bar{w}_k^T \bar{\phi}(\bar{x}_n)$$

(3) Optimization :

Softmax derivative : $\frac{\partial}{\partial x}\left(\frac{e^x}{e^x + e^y}\right)$

Recall : $\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{u'v - v'u}{v^2}$

Thus : $\frac{\partial}{\partial x}\left(\frac{e^x}{e^x+e^y}\right) = \frac{e^x(e^x+e^y) - e^{2x}}{(e^x+e^y)^2} = \frac{e^x(e^x+e^y-e^x)}{(e^x+e^y)^2}$

$$= \frac{e^x}{(e^x+e^y)}\left[\frac{e^x+e^y}{(e^x+e^y)} - \frac{e^x}{(e^x+e^y)}\right] = y_{nk}(1 - y_{nk})$$

$$\frac{\partial}{\partial \bar{w}_j} E(\bar{w}_1 \dots \bar{w}_K) = \frac{\partial}{\partial \bar{w}_j} \sum_n \sum_k t_{nk} \log y_{nk}$$

$\bar{w}_j$: M×1

$$= -\sum_n t_{nj} \left[\frac{1}{y_{nj}} \left(y_{nj}(1 - y_{nj}) \bar{\phi}_n \right) \right]$$

$\bar{\phi}_n$: M×1

$$= \sum_n (y_{nj} - t_{nj}) \bar{\phi}_n$$

$$= \Phi^T \bar{R}_j$$

Φ: M×N, $\bar{R}_j$: N×1

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_j \exp(a_{nj})}$$

$$\bar{R}_j = \begin{bmatrix} y_{1j} - t_{1j} \\ \vdots \\ y_{Nj} - t_{Nj} \end{bmatrix}$$

y_{ij}: unknown; t_{ij}: known

$$\frac{\partial}{\partial W} E(W) = \Phi^T R$$

W: M×K, Φ: M×N, R: N×K

$$R = \begin{bmatrix} \bar{R}_1 & \bar{R}_2 & \dots & \bar{R}_K \end{bmatrix}$$

Solvable using Newton's method (see textbook p. 209)

Q : What is the relationship between sigmoid & softmax ?

A : Binary classes : sigmoid

Bayesian perspective :

$$p(t_n = 1 \mid \bar{\omega}) = \frac{p(\bar{\omega} \mid t_n = 1)\, p(t_n = 1)}{p(\bar{\omega})}$$

$$= \frac{p(\bar{\omega} \mid t_n = 1)\, p(t_n = 1)}{\sum_{t_n = \{0,1\}} p(\bar{\omega} \mid t_n)\, p(t_n)}$$

$$= \frac{p(\bar{\omega} \mid t_n = 1)\, p(t_n = 1)}{p(\bar{\omega} \mid t_n = 0)\, p(t_n = 0) + p(\bar{\omega} \mid t_n = 1)\, p(t_n = 1)}$$

$$= \frac{1}{1 + \dfrac{p(\bar{\omega} \mid t_n = 0)\, p(t_n = 0)}{p(\bar{\omega} \mid t_n = 1)\, p(t_n = 1)}} = \sigma(a_n)$$

Multi class : softmax

$$p(t_{n,k} \mid W) = \frac{\exp(a_{n,k})}{\sum_j \exp(a_{n,j})}$$

Consider $K = 2$:

$$p(t_{n,k} = 1 \mid W) = \frac{\exp(a_{n,k=1})}{\exp(a_{n,k=0}) + \exp(a_{n,k=1})} = \frac{1}{1 + \dfrac{\exp(a_{n,k=0})}{\exp(a_{n,k=1})}}$$

Recall : in softmax, we define :

$$\exp(a_{n,k}) = p(W \mid t_n = C_k)\, p(t_n = C_k)$$

$$p(t_{n,k} = 1 \mid W) = \frac{1}{1 + \dfrac{p(W \mid t_n = 0)\, p(t_n = 0)}{p(W \mid t_n = 1)\, p(t_n = 1)}} = \sigma(a_n)$$

Softmax for two classes is the same as sigmoid.

PROBABILISTIC GENERATIVE MODELS

Assuming two classes: $t_n \in \{C_1, C_2\}$ e.g. $C_1 = 1$, $C_2 = 0$

$$p(t_n = C_1) = \vartheta \quad ; \quad \vartheta = \text{unknown}$$

$$p(t_n = C_2) = 1 - \vartheta$$

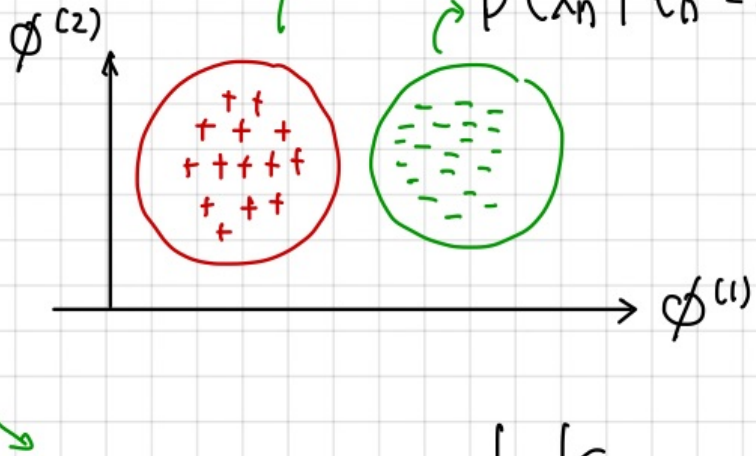
$$p(\bar{x}_n, t_n = C_1) = p(\bar{x}_n \mid t_n = C_1)\, p(t_n = C_1)$$

$$p(\bar{x}_n, t_n = C_2) = p(\bar{x}_n \mid t_n = C_2)\, p(t_n = C_2)$$

where:

$$p(\bar{x}_n \mid t_n = C_1) = G(\bar{\phi}(\bar{x}_n); \bar{\mu}_1, \Sigma)$$

$$p(\bar{x}_n \mid t_n = C_2) = G(\bar{\phi}(\bar{x}_n); \bar{\mu}_2, \Sigma)$$



$$p(\bar{x}_n, t_n = C_1) = \vartheta\, G(\bar{\phi}(\bar{x}_n); \bar{\mu}_1, \Sigma)$$

$$p(\bar{x}_n, t_n = C_2) = (1 - \vartheta)\, G(\bar{\phi}(\bar{x}_n); \bar{\mu}_2, \Sigma)$$

Assumption here: $\Sigma_1 = \Sigma_2 = \Sigma$ for simplicity discussion

Our goal is to estimate: $\bar{\mu}_1, \bar{\mu}_2, \Sigma$ and π

Hence, the likelihood with respect to the unknowns:

$$\underset{N\times 1}{p(\bar{t}, \underset{N\times D}{X} \mid \pi, \underset{1\times 1}{\bar{\mu}_1}, \underset{M\times 1}{\bar{\mu}_2}, \underset{M\times 1}{\Sigma}, \underset{M\times M}{})} = \prod_n \left[\pi G(\underset{M\times 1}{\bar{\phi}(x_n)}; \underset{M\times 1}{\bar{\mu}_1}, \Sigma)\right]^{t_n} \left[(1-\pi) G(\bar{\phi}(x_n); \bar{\mu}_2, \Sigma)\right]^{1-t_n}$$

By applying the negative log, the terms depending on θ:

$$E(\pi) = \sum_n \left(t_n \log \pi + (1-t_n) \log(1-\pi)\right)$$

$$\frac{\partial E(\pi)}{\partial \pi} = \sum_n t_n \frac{1}{\pi} + (1-t_n) \frac{1}{(1-\pi)} (-1) = 0$$

$$= \frac{1}{\pi} \sum_n t_n - \frac{1}{(1-\pi)} \sum_n (1-t_n) = 0$$

$$\frac{1}{\pi} \sum_n t_n = \frac{1}{(1-\pi)} \sum_n (1-t_n) \rightarrow \frac{(1-\pi)}{\pi} = \frac{N - \sum t_n}{\sum t_n}$$

$$\frac{1}{\pi} - 1 = \frac{N}{\sum t_n} - 1 \Rightarrow \boxed{\pi = \frac{\sum t_n}{N} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}}$$

[•] Terms depending on μ_1:

$$E(\bar{\mu}_1) = -\sum_n t_n \log\left(\frac{1}{(2\pi)^{M/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\bar{\phi}-\bar{\mu}_1)^T \Sigma^{-1}(\bar{\phi}-\mu_1)\right)\right)$$

$$= \frac{1}{2}\sum_n t_n (\bar{\phi}-\mu_1)^T \Sigma^{-1}(\phi-\mu_1) + \text{const.}$$

Minimization:

$$(M\times 1)\quad \frac{\partial E(\bar{\mu}_1)}{\partial \bar{\mu}_1} = \sum_n t_n \underset{M\times M}{\Sigma^{-1}} \underset{M\times 1}{(\bar{\phi}-\bar{\mu}_1)} \underset{M\times 1}{} = 0$$

$$\sum_n t_n \Sigma^{-1}\bar{\mu}_1 = \sum_n t_n \Sigma^{-1}\bar{\phi}(x_n)$$

$$\boxed{\bar{\mu}_1 = \frac{1}{N_1}\sum_n t_n \bar{\phi}(x_n)} \xrightarrow{\text{Consequently}} \boxed{\bar{\mu}_2 = \frac{1}{N_2}\sum_n (1-t_n)\bar{\phi}(x_n)}$$

[•] Terms depending on Σ:

$$E(\Sigma) = \frac{1}{2}\sum_n t_n \log|\Sigma| + \frac{1}{2}\sum_n t_n(\bar{\phi}(x_n)-\bar{\mu}_1)^T\Sigma^{-1}(\bar{\phi}(x_n)-\bar{\mu}_1)$$
$$+ \frac{1}{2}\sum_n (1-t_n)\log|\Sigma| + \frac{1}{2}\sum_n (1-t_n)(\bar{\phi}(x_n)-\bar{\mu}_2)^T\Sigma^{-1}(\bar{\phi}(x_n)-\bar{\mu}_2)$$

$$= \frac{N}{2}\log|\Sigma| + \frac{N}{2}\text{Tr}\left(\Sigma^{-1}\underset{M\times M}{\mathbb{S}}\right)$$

where:

$$\mathbb{S} = \frac{N_1}{N}\mathbb{S}_1 + \frac{N_2}{N}\mathbb{S}_2$$

$$\mathbb{S}_1 = \frac{1}{N_1}\sum_n (\bar{\phi}(x_n)-\bar{\mu}_1)(\bar{\phi}(x_n)-\bar{\mu}_1)^T$$

$$\mathbb{S}_2 = \frac{1}{N_2}\sum_n (\bar{\phi}(x_n)-\bar{\mu}_2)(\bar{\phi}(x_n)-\bar{\mu}_2)^T$$

Minimization:

$$\frac{\partial E(\Sigma)}{\partial \Sigma} = 0 \quad \rightarrow \quad \boxed{\Sigma = \mathbb{S}}$$

$$E(\Sigma) = \frac{N}{2}\log|\Sigma| + \frac{N}{2}\mathrm{Tr}\left(\Sigma^{-1}\underset{M\times M}{S}\right)$$

$$\frac{\partial}{\partial\Sigma}E(\Sigma) = \underbrace{\frac{\partial}{\partial\Sigma}\left(\frac{N}{2}\log|\Sigma|\right)}_{①} + \underbrace{\frac{\partial}{\partial\Sigma}\left(\frac{N}{2}\mathrm{Tr}\left(\Sigma^{-1}S\right)\right)}_{②} = 0$$

(1) $\frac{N}{2}\frac{\partial}{\partial\Sigma}\log|\Sigma|$ $\qquad \frac{d}{dx}|A(x)| = \frac{d}{dx}|A| = |A|\,\mathrm{Tr}\left(A^{-1}\frac{d}{dx}A\right)$

$$= \frac{N}{2}\frac{1}{|\Sigma|}\frac{\partial}{\partial\Sigma}|\Sigma| = \frac{N}{2}\frac{1}{|\Sigma|}|\Sigma|\,\mathrm{Tr}\left(\Sigma^{-1}\right) = \frac{N}{2}\mathrm{Tr}\left(\Sigma^{-1}\right)$$

(2) $\frac{N}{2}\frac{\partial}{\partial\Sigma}\mathrm{Tr}\left(-\Sigma^{-1}S\right) = -\frac{N}{2}\mathrm{Tr}\left(\left(\Sigma^{-1}\right)^2 S\right)$

Hence: $\frac{\partial}{\partial\Sigma}E(\Sigma) = \mathrm{Tr}\left(\Sigma^{-1}\right) - \mathrm{Tr}\left(\left(\Sigma^{-1}\right)^2 S\right) = 0$

$$\mathrm{Tr}\left(\left(\Sigma^{-1}\right)^2 S\right) = \mathrm{Tr}\left(\Sigma^{-1}\right) \longrightarrow \boxed{\Sigma = S}$$